International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 6, Issue 5, Aug – Sep 2017; 7-18
© IASET

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

STATISTICAL PATTERN CLASSIFICATION OF DIRECT SPEECHES IN CHILDREN STORIES

MENAKA SIKDAR[1] & PRANITA SARMAH[2]

[1]Research Scholar, Department of Statistics, Gauhati University, Guwahati, Assam, India

[2]Professor, Department of Statistics, Gauhati University, Guwahati, Assam, India

ABSTRACT

This paper presents a study of three languages, namely Assamese, Bengali and English. The main objective of this study is to pattern classification of direct speeches with special reference to children's stories, in order to find the distinction among all these languages. We consider only the children's stories because, they are found to be similar all over the world with different flavours produced by different cultures, languages and time. We have taken 30 Assamese stories from 'Burhi Aai'r Xaadhu'(literary translated to Grandma's tales), 27 Bengali stories from "Tuntunir Boi" (Book of the tailor-bird), 62 English stories from Grimm's fairy tales and 16 English stories from Anderson's fairy tales for collecting data. Detailed statistical analyses have been performed by quantifying the direct speeches and presenting them graphically. Non-parametric approaches have been used to test the significant differences among the direct speeches under consideration. It has been shown that there exist significant differences among the direct speeches written by different authors in different languages. Kolmogorov Goodness –of- Fit test, Kruskal-Wallis test and Squared Ranks Test were used for this purpose.

KEYWORDS: Empirical Distribution, Kolmogorov Goodness –of- Fit Test Kruskal-Wallis Test, Non-Parametric Tests, Squared Ranks Test

1. INTRODUCTION

Children's stories are written mainly to entertain the children. But, a careful observation on these stories reveals that besides entertainment these stories are sources of knowledge and values that are important for living the life. It helps the children to develop imagination, language skills, cooperation and other social skills such as confidence, and creative expression. The narrators (authors) not only tell the audience (the kids) what happened, but also try to show them what occurred. The authors describe the scenes of the stories to the young audience and help them to visualize different characters in a story with the help of **direct speeches**. Direct speech refers to the quoted words of a character given by the narrators. The author tries to represent the words of the characters what they exactly say. Almost all the authors of the children stories use more direct speech to represent the emotions, moods, natures and tones of the different characters in a unique manner to make it more and more interesting. The direct speeches are used to present the arguments, quarrels and conversation among the different characters which provide excitement, tension, and amusement to the readers.

Quite a number of research articles are available in literature on direct speech, which serves specific goals in communication. Clark and Gerrig (1990) mentioned that important function of direct speech is to provide a demonstration of speech, whereas indirect speech delivers a description of what was said. The importance of direct speech used in literature in a unique manner to make it more and more interesting, when compared to indirect speech, have been discussed by (Wierzbicka, 1974; Li, 1986; Tannen, 1989; Yao, Belin & Scheepers, 2011). Direct speech is usually more vivid than indirect speech; it is often used at the climax of stories, and is proposed to be an effective way of conveying the point of a narrative (Mayes, 1990). A study was conducted by (Groenewold et al, 2015) to examine whether the effects of direct and indirect speech constructions on discourse comprehension are the same or different for the two languages viz Dutch and English. Groenewold et al. (2014b) showed that, for Dutch individuals with and without aphasia, narratives containing direct speech constructions were better comprehended than narratives with indirect speech constructions. Sikdar and Sarmah (2017) presented a study in three languages, namely Assamese, Bengali and English for pattern recognition of language model with special reference to children's stories. Non-parametric approaches have been used to test the significant differences among the texts under consideration.

However, in our present article, we are mainly concerned with the statistical pattern classification of direct speeches in children's stories. In this article, we have considered the following variables that will actually help us to recognize the pattern of the direct speeches in children stories- (i) **the total number of Direct speeches, (ii) total number of sentences within Direct speeches (iii) maximum number of sentences within the direct speech (iv) the ratio of total number of sentences within the direct speeches to the total number of sentences in the story (v) total number of words within Direct speeches and (vi) maximum number of words within the direct speech of different Stories under different languages.**

2. OBJECTIVES OF THE STUDY

The main objective of our study is to recognize patterns of direct speeches in children's stories written by several authors in different languages. Moreover, our aim is to answer the research questions mentioned below. For analysis, we have considered stories under different languages, namely **Assamese, Bengali, English1 (Grimm's' Fairy tales) and English2 (Andersen's Fairy tales).**

(a) Whether the distributions of the random variables (i), (ii), (iii), (iv), (v) and (vi) (as mentioned in **section 1**) have come from normal population.

(b) Whether there exists a significant difference between the distribution functions of the random variables (i), (ii), (iii), (iv), (v) and (vi) under different languages.

3. SOURCES OF DATA

For analyzing the direct speeches of children's stories, the data have been collected from the stories by **Sahityo –rothi Laxminath Bezbarua,** Upendrakishore Roy Choudhury, the Grimm brothers and Hans Christian Andersen. Thirty Assamese stories from 'Burhi Aai'r Xaadhu' (literary translated to Grandma's Tales), twenty seven Bengali stories from 'Tuntunir Boi'(Book of the tailor-bird) , sixty two English stories from Grimm's fairy tales and sixteen English stories from Andersen's fairy tales have been selected for the purpose of analysis.

4. MATERIALS AND METHODS

The concept of Empirical distribution function is used for studying the probabilistic structure of the variables (i),(ii),(iii),(iv),(v) and (vi) (as mentioned in section 1) under different languages. Run test based on median has been conducted for the distributions of (i), (ii),(iii),(iv),(v) and (vi) and found that these distributions are random under different languages. Non parametric techniques are used for analysis of language pattern of direct speeches under consideration.

4.1 Some Definitions

Direct or **quoted speech** is spoken or written text that reports speech or thought in its original form phrased by the original speaker; in narrative, it is usually enclosed in quotation marks. $D_d(k,l)$ be the d^{th} **Direct Speech** in k^{th} story under l^{th} language, $\forall d = 1,2,\dots,\delta$, k=1,2,..,r, l=1,2,…,s

The following statistics are defined for analyzing **the** direct speeches.

(i) N_D (k, l) be the total number of speeches enclosed in quotation marks with respect to the k^{th} story under l^{th} language.

(ii) S_d (k,l) be the **total number sentences within the d^{th} direct speech** in kth story under lth language. $\forall d = 1,2,\dots,\delta$, k=1,2,..,r, l=1,2,…,s.

$S_D(k,l) = \sum_{d=1}^{\delta} S_d\,(k,l)$ be the **total number of sentences within the direct speeches** in kth story under lth language.

(iii) Let $S_{(1)}(k,l) \leq S_{(2)}(k,l) \leq \cdots \leq S_{(\delta)}(k,l)$ be the ordered statistics obtained from S_d(k,l), $\forall d = 1,2,\dots,\delta$,

Maximum number of sentences within the direct speech in kth story under lth language i.e Max S_d(k,l)= $S_{(\delta)}(k,l)$, d=1,2,.., δ

(iv) $\mathbf{S_{kl}}$ **is the total number sentences** in the kth story under lth language.

$R_{DS}(k,l) = \frac{S_D(k,l)}{S_{kl}}$ be **the ratio of total number of sentences within the direct speeches to the total number of sentences** in the kth story under lth languages.

(v) $w_d(k,l)$ be **the total number of words within dth direct speech** in kth story under lth languages, $\forall d = 1,2,\dots,\delta$, k=1,2,..,r, l=1,2,…,s .

$w_D(k,l) = \sum_{d=1}^{\delta} w_d(k,l)$ be **the total number of words within the direct speeches** in kth story under lth language.

(vi) Let $w_{(1)}(k,l) \leq w_{(2)}(k,l) \leq \cdots \leq w_{(\delta)}(k,l)$ be the ordered statistics obtained from $w_d(k,l)$ $\forall d = 1,2,\dots,\delta$.
Maximum number of words within the direct speech in kth story under lth language i.e Max $w_d(k,l) = w_{(\delta)}(k,l)$, d=1, 2,.., δ d

4.2 SOME STATISTICAL TOOLS FOR ANALYZING THE PATTERNS OF THE DIRECT SPEECHES

4.2.1 Empirical Distribution Function

The Empirical Distribution Functions are used to study the probabilistic structures of the random variables under consideration. In case of the distribution of total number of direct speeches of different stories under lth language, data consist of a random sample $N_D(1,l), N_D(2,l), \dots, N_D(r,l)$ **of size r.**

The empirical distribution function under lth language,

$F_{N_Dl}(x) = (number\ of\ N_D(k,l) \leq x)/r$ Where k=1, 2, r, l=1, 2, 3, 4

Similarly, we can obtain the empirical distribution functions of the distributions, namely (ii), (iii), (iv), (v) and (vi) respectively.

4.2.2 The Kolmogorov Goodness –of- Fit Test

The Kolmogorov Goodness –of- Fit test has been adopted to test the normality of the distributions under study. ***In case of the distribution of total number*** **of direct speeches of different stories under lth language,** the data consists of a random sample $N_D(1,\ l),N_D(2,l),\ldots,N_D(r,l)$ **of size r** associated with some unknown distribution function, denoted by $Q(N_D l)$.

Test Statistic

Let $F(N_D l)$ be the empirical distribution function based on the random sample $N_D(1,l),\ N_D(2,l),\ldots,N_D(r,l)$. Let $Q^*(N_D l)$ be a completely specified hypothesized distribution function which is considered here as a **normal probability distribution function**.

Let the test statistic T be the greatest vertical distance between $F\ (N_D l)$ and $Q^*(N_D l)$. Mathematically

$$T=\sup_{N_D l} |Q^*(\boldsymbol{N_D}l) - F(\boldsymbol{N_D}l)|$$

Null Distribution: when Q(N_Dl) is continuous and the null hypothesis is true, the approximate distribution function of T is $p(T \leq N_D l) = [G(N_D l)]^2$, where

$$G(N_D l) = 1 - N_D l \sum_{p=0}^{[r(1-N_D l)]} \binom{r}{p}\left(1 - N_D l - \frac{p}{r}\right)^{r-p} \left(N_D l + \frac{p}{r}\right)^{p-1}$$

Where $[r\ (1\text{-}N_D l)]$ is the greatest integer less than or equal to $r(1\text{-}N_D l)$.

Hypotheses

(Two-sided Test) The null hypothesis is to be tested

H$_0$: $Q(N_D l)=Q^*(N_D l)$ for all $N_D l$ from $-\infty$ to $+\infty$

H$_1$: $Q(N_D l)\neq Q^*(N_D l)$ for at least one value of $N_D l$

Similar procedure and hypotheses are used for the distributions namely (ii), (iii), (iv), (v) and (vi) respectively.

4.3.3 The Kruskal-Wallis Test

The Kruskal-Wallis test is applied to compare the means of the distributions coming from different languages **for the distributions of total number of Direct speeches of different stories*;*** our data consists of 4 random samples of different sizes. Let r_1, r_2, r_3 and r_4 are the sample sizes (number of stories) of Assamese, Bengali, English1 and English2 respectively. The data may be arranged as below:

Table 1

Sample1(Assamese)	Sample2(Bengali)	Sample3(English1)	Sample4(English2)
$N_D(1,1)$	N_D (1,2)	N_D (1,3)	N_D (1,4)
N_D (2,1)	N_D (2,2)	N_D (2,3)	N_D (2,4)
…	…	…	…
N_D (r_1 ,1)	N_D (r_2 ,2)	N_D (r_3 ,3)	N_D (r_4 ,4)

Let r be the total number of observations (stories). $r = \sum_{l=1}^{4} r_l$

Let ρ[N_D(k, l)] be the rank assigned to N_D(k,l) .Let ρ_l be the sum of ranks assigned to the lth sample. $\rho_l = \sum_{k=1}^{r_l} \rho[N_D(k,l)]$ l=1, 2, 3, 4

We assign the average rank to each of the tied observations.

Test Statistic

The test statistics T is defined as

$T = \frac{1}{c^2}\left(\sum_{l=1}^{4} \frac{\rho_l{}^2}{r_l} - \frac{r(r+1)^2}{4}\right)$ Where $c^2 = \frac{1}{r-1}\left(\sum_{allranks} \rho[N_D(k,l)]^2 - \frac{r(r+1)^2}{4}\right)$

If there are no ties c^2 simplifies to r (r+1)/12 and test statistics reduces to $T = \frac{12}{r(r+1)}\sum_{l=1}^{4}\frac{\rho_l{}^2}{r_l} - 3(r+1)$

Null Distribution

The chi-square distribution with 4-1=3 degrees of freedom is used as an approximation for the null distribution of T.

Hypotheses

$\boldsymbol{H_{01}}$**:** All of the four population distribution functions of **total number of Direct speeches** of different stories under different languages are identical.

$\boldsymbol{H_{11}}$**:** The four populations of **total number of Direct speeches** of different stories under different languages differ significantly corresponding to their means.

Multiple Comparisons

When the null hypothesis is rejected, we may use the following procedure to determine which pairs of populations tend to differ. The populations say l and k seem to be different, if the following inequality is satisfied: $\left|\frac{\rho_l}{r_l} - \frac{\rho_k}{r_k}\right| > t_{1-(\alpha/2)}\left(c^2\frac{r-1-T}{r-s}\right)^{\frac{1}{2}}\left(\frac{1}{r_l}+\frac{1}{r_k}\right)^{\frac{1}{2}}$

Where, s is the number of random samples, ρ_l and ρ_k is the rank sums of the two samples, $t_{1-(\alpha/2)}$ is the $[1-(\alpha/2)]$ quantile of the t distribution with r-s degrees of freedom.

Similar procedure and hypotheses are adopted for the distributions, namely (ii), (iii), (iv), (v) and (vi), respectively.

4.3.4 The Squared Ranks Test

The squared rank test is used to compare the variances of the populations. For the distributions of total number of

Direct speeches of different stories, data consists of 4 independent samples which are given in **Section 4.3.3.** For this analysis, we subtract the population mean from each observation (or its sample mean when population mean is unknown) and convert these differences to absolute differences. Then, we rank the combined absolute differences from smallest to largest, assigning average ranks in case of ties. Then, we compute the sum of squares of the ranks of each sample (language).

Test Statistic

The test statistics is $T = \frac{1}{D^2}\left[\sum_{l=1}^{4} \frac{Q_l^2}{r_l} - r(\bar{Q})^2\right]$ .Where r_l= number of observations in lth sample. r=$\sum_{l=1}^{4} r_l$

Let, Q_l= the sum of the squared ranks in the lth sample, l=1, 2, 3, 4.

$\bar{Q} = \frac{1}{r}\sum_{l=1}^{4} Q_l$ and $D^2 = \frac{1}{r-1}[\sum_{k=1}^{r} \rho_k^4 - r(\bar{Q})^2]$

and let $\sum_{k=1}^{r} \rho_k^4$ represents the sum resulting after raising each rank to the fourth power.

If there is no ties, D^2 = r(r+1)(2r+1)(8r+11)/180 and $\bar{Q}$=(r+1)(2r+1)/6

Null Distribution

The null distribution of T is approximately the chi- squared distribution with 4-1=3 degrees of freedom.

Hypotheses

$\boldsymbol{H_{01}}$**:** All of the four populations of **total number of direct speeches** contained in different stories under different languages are identical, except for possibly different means.

$\boldsymbol{H_{11}}$**:** The four populations of **total number of direct speeches words** contained in different stories under different languages do not have identical variance.

Multiple Comparisons

When the null hypothesis is rejected, we may use the following procedure to determine which pairs of populations tend to differ. The populations say l and k seem to be different if the following inequality is satisfied: $\left|\frac{Q_l}{r_l} - \frac{Q_k}{r_k}\right| > t_{1-(\alpha/2)} \left(D^2 \frac{r-1-T}{r-s}\right)^{\frac{1}{2}} \left(\frac{1}{r_l} + \frac{1}{r_k}\right)^{\frac{1}{2}}$

Where, s be the number of random samples, Q_l and Q_k are the sums of the squared ranks in the lth and kth samples, respectively, $t_{1-(\alpha/2)}$ is the[$1 - (\alpha/2)$] quantile of the t distribution with r-s degrees of freedom.

Similar procedure and hypotheses are adopted for the distributions, namely (ii), (iii), (iv), (v) and (vi), respectively.

5. STATISTICAL ANALYSES AND RESULTS

The empirical distribution functions of the random variables mentioned above are presented below.

[Assamese → 🟥*,Bengali →*🟦 *,English1 →*🟨 *,English2 →*⬛ *]*

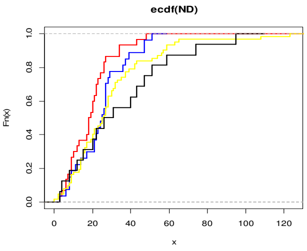


Figure 1

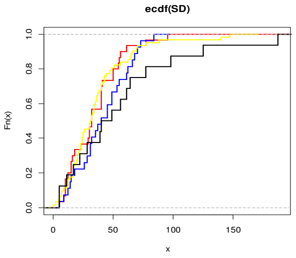


Figure 2

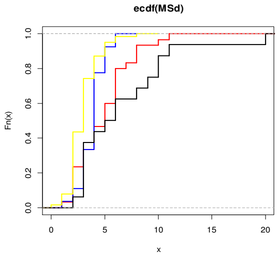


Figure 3

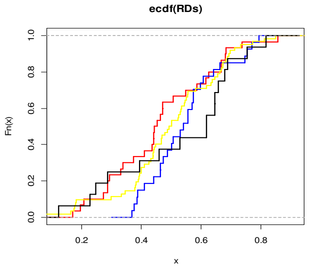


Figure 4

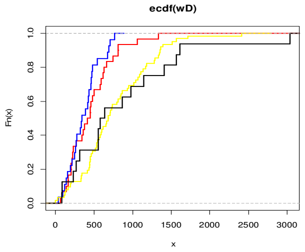
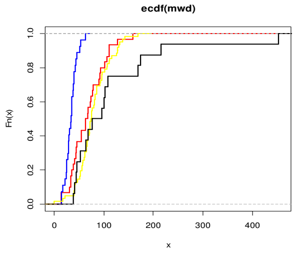


Figure 5



Figures 6

[The empirical distribution functions of the distributions of (i),(ii),(iii),(iv),(v) and (vi) under Assamese, Bengali, English1 and English2 are represented in Figure 1,2,3,4,5and 6 respectively]

Results of Kolmogorov Goodness –of- Fit test are obtained by using **SPSS soft-ware** and are given in the following table.

Table 2: Results of Kolmogorov Goodness –of- Fit Test

Language	Characteristics	Distributions of the Random Variables					
		(i)	(ii)	(iii)	(iv)	(v)	(vi)
Assamese	sample size	30	30	30	30	30	30
	Mean	18.93	34.23	4.87	0.4648	455.27	67.50
	Standard deviation	11.033	21.574	2.515	0.1740	295.219	34.510
	absolute difference	0.099	0.110	0.126	0.119	0.106	0.109
	Positive difference	0.099	0.110	0.126	0.119	0.106	0.109
	Negative difference	-0.074	-0.088	-0.094	-0.077	-0.092	-0.072
	K.S test statistic ,*Z*	0.542	0.604	0.691	0.653	0.583	0.600
	p-value	**0.930**	**0.859**	**0.726**	**0.778**	**0.886**	**0.865**
Bengali	sample size	27	27	27	27	27	27
	Mean	24.44	41.22	3.81	0.5453	368.26	33.63
	Standard deviation	12.777	21.689	1.145	0.1229	196.865	11.666
	absolute difference	0.138	0.080	0.231	0.111	0.102	0.083
	Positive difference	0.138	0.080	0.214	0.111	0.102	0.083
	Negative difference	-0.097	-0.078	-0.231	-0.096	-0.090	-0.056
	K.S test statistic ,*Z*	0.720	0.416	1.2	0.579	0.528	0.431
	p-value	**0.678**	**0.995**	**0.112**	**0.891**	**0.943**	**0.992**
English 1	sample size	62	62	62	62	62	62
	Mean	29.13	35.73	2.94	0.4961	768.76	78.39
	Standard deviation	22.69	27.852	1.389	0.1706	466.710	32.135
	absolute difference	0.159	0.157	0.223	0.099	0.106	0.112
	Positive difference	0.159	0.157	0.223	0.064	0.106	0.112
	Negative difference	-0.100	-0.100	-0.170	-0.099	-0.050	-0.068
	K.S test statistic ,*Z*	1.255	1.239	1.759	0.762	0.835	0.882
	p-value	**0.086**	**0.093**	**0.004**	**0.574**	**0.488**	**0.418**

Table 2: Contd.,

Language	Characteristics	Distributions of the Random Variables					
		(i)	(ii)	(iii)	(iv)	(v)	(vi)
English2	sample size	16	16	16	16	16	16
	Mean	34.38	55.25	6.62	0.5233	870.88	115.38
	Standard deviation	26.22	48.641	4.66	0.2094	764.176	104.336
	absolute difference	0.125	0.179	0.178	0.233	0.181	0.278
	Positive difference	0.125	0.179	0.178	0.121	0.181	0.278
	Negative difference	-0.116	-0.151	-0.160	-0.233	-0.150	-0.229
	K.S test statistic ,*Z*	0.501	0.714	0.713	0.934	0.723	1.113
	p-value	**0.963**	**0.687**	**0.689**	**0.348**	**0.673**	**0.168**

Conclusions: From table 2, it has been noticed that the p-values of the test statistics for the distributions of (i),(ii), (iii)(except for the stories under English1), (iv), (v) and (vi) under different languages are greater than 0.05.Therefore, we may accept our null hypotheses at the 5 % level of significance and may conclude that the distributions of these random variables under different languages namely Assamese, Bengali, **English1** and English2 are **normally distributed**. The p-value of the test statistic for the distribution of (iii) under English1 stories is 0.004 which is less than 0.05. Therefore, we may reject the null hypothesis at the 5 % level of significance and may conclude that the distribution of random variable (iii) under **English1**is not **normally distributed**. The results of the *Kruskal-Wallis tests* are obtained by using **SPSS software** and are given in **the following table.**

Table 3: Results of Kruskal-Wallis Test

Distribution	Language	Sample Size	Mean Rank	Test Statistic (Chi-Square)	D.F	P-Value
(i)	Assamese	30	52.43	6.693	3	0.082
	Bengali	27	69.02			
	English1	62	72.40			
	English2	16	78.41			
(ii)	Assamese	30	63.70	3.993	3	0.262
	Bengali	27	76.70			
	English1	62	63.31			
	English2	16	79.56			
(iii)	Assamese	30	84.07	26.727	3	0.000
	Bengali	27	75.39			
	English1	62	50.30			
	English2	16	94.00			
(iv)	Assamese	30	58.37	3.733	3	0.292
	Bengali	27	76.65			
	English1	62	67.02			
	English2	16	75.28			
(v)	Assamese	30	54.17	22.374	3	0.000
	Bengali	27	44.81			
	English1	62	81.87			
	English2	16	79.31			
(vi)	Assamese	30	67.83	45.005	3	0.000
	Bengali	27	24.70			
	English1	62	81.69			
	English2	16	88.31			

Conclusion: From table 3, it has been noticed that the p-values of the test statistics of the distributions of (i),(ii) and (iv) under the 4 different languages are greater than 0.05. Therefore, we may accept our null hypotheses at the 5 %

level of significance and may conclude that these distributions of different stories under 4 different languages may have identical mean i.e. they are not significantly different corresponds to their mean. On the other hand, the p-values of the test statistics for **the distributions of (iii), (v) and (vi)** under 4 different languages are less than 0.05. Therefore, we may reject our null hypotheses at the 5 % level of significance and may conclude that **these distributions** under 4 different languages do not have identical means i.e. they are all significantly different corresponding to their means. However, when such a null hypothesis is rejected, it is a normal practice to perform a multiple comparison procedure to determine which pairs of population tend to differ. Calculations for multiple comparisons are given in **the following table**

Table 4: Results of Multiple Comparisons under Kruskal-Wallis Test

Numerical Data	Languages	$\left\|\frac{\rho_l}{r_l} - \frac{\rho_k}{r_k}\right\|$	$t_{1-(\alpha/2)}\left(c^2\frac{r-1-T}{r-s}\right)^{\frac{1}{2}}\left(\frac{1}{r_l}+\frac{1}{r_k}\right)^{\frac{1}{2}}$	Results
(iii)Maximum number of sentences within the direct speech	1 and 2	8.68	18.4038	Not significantly different
	1and 3	33.77	15.4294	significantly different
	1 and 4	9.93	21.4768	Not significantly different
	2 and 3	25.09	15.9966	significantly different
	2 and 4	18.61	21.8879	Not significantly different
	3 and 4	43.7	19.4573	significantly different
(v)Total number of words within Direct speeches	1 and 2	9.36	18.7734	Not significantly different
	1and 3	27.7	15.7393	Significantly different
	1 and 4	25.14	21.9082	Significantly different
	2 and 3	37.06	16.3180	Significantly different
	2 and 4	34.5	22.3276	Significantly different
	3 and 4	2.56	19.8445	Not significantly different
(vi)Maximum number of words within the direct speech	1 and 2	43.13	16.7627	significantly different
	1and 3	13.86	14.0536	Not significantly different
	1 and 4	20.48	19.5617	significantly different
	2 and 3	56.99	14.5702	significantly different
	2 and 4	63.61	19.9362	Significantly different
	3 and 4	6.62	17.7191	Not significantly different

[Language takes value 1,2,3,4 for Assamese, Bengali, English1 and English2 stories respectively]

Results of the Squared Rank Test are given in the following table.

Table 5: Results of the Squared Rank Test

Numerical Data	Language	Sample Size (R_l)	The Sum of the Squared Ranks, Q_l	Test Statistic (Chi-Square)	D.F.
(i)	Assamese	30	111264.5	15.904	3
	Bengali	27	128304.5		
	English1	62	434308		
	English2	16	155345.5		
(ii)	Assamese	30	157364.5	8.9465	3
	Bengali	27	164334.5		
	English1	62	348732.5		
	English2	16	158803.5		
(iii)	Assamese	30	257198	40.9277	3
	Bengali	27	100047.5		
	English1	62	267154		
	English2	16	203121		

Table 5: Contd.,

Numerical Data	Language	Sample Size (R_l)	The Sum of the Squared Ranks, Q_l	Test Statistic (Chi-Square)	D.F.
(iv)	Assamese	30	203007	6.2022	3
	Bengali	27	117362		
	English1	62	373224.5		
	English2	16	135662.5		
(v)	Assamese	30	129426	26.4945	3
	Bengali	27	72805		
	English1	62	463175		
	English2	16	163854		
(vi)	Assamese	30	210206.5	36.9031	3
	Bengali	27	49983		
	English1	62	373415.5		
	English2	16	195634		

Conclusion: The critical value of $\chi^2_{(3),0.95}$ is 7.815. From **Table 6**, it has been noticed that the calculated values of the test statistics under the distributions of **(I),(ii),(iii),(v) and (vi)** under 4 different languages are greater than 7.815. Therefore, we may reject our null hypotheses at the 5 % level of significance and may conclude that the distributions of **these random variables** of different stories under 4 different languages do not have identical variance i.e. they are significantly different corresponds to their variance. However, the calculated value of the test statistic under the distribution of (iv) under 4 different languages is less than 7.815. Therefore, we may accept our null hypothesis at the 5 % level of significance and may conclude that the distributions of (iv) under different languages have identical variances i.e. they are not significantly different corresponds to their variances. However, when such a null hypothesis is rejected, it is a normal practice to perform **Multiple Comparisons Procedure** to determine which pairs of populations tend to differ. Calculations for multiple comparisons are given in **the following table**

Table 6: Results of Multiple Comparisons under the Squared Rank Test

Numerical Data	Languages	$\left\lvert\frac{Q_l}{r_l}-\frac{Q_k}{r_k}\right\rvert$	$t_{1-(\alpha/2)}\left(D^2\frac{r-1-T}{r-s}\right)^{\frac{1}{2}}\left(\frac{1}{r_l}+\frac{1}{r_k}\right)^{\frac{1}{2}}$	Inference
(i)Total number of Direct speeches	1 and 2	1043.2	2710.786	Not significantly different
	1and 3	3296.15	2272.677	significantly different
	1 and 4	6000.28	3163.433	significantly different
	2 and 3	2252.95	2356.23	Not significantly different
	2 and 4	4957.08	3223.983	Significantly different
	3 and 4	2704.13	2865.443	Not significantly different
(ii)Total number of sentences within Direct speeches	1 and 2	840.9796	2789.654	Not significantly different
	1and 3	379.2344	2338.799	significantly different
	1 and 4	4679.7354	3255.47	significantly different
	2 and 3	461.7452	2424.782	significantly different
	2 and 4	3838.7558	3317.781	Significantly different
	3 and 4	4300.501	2948.81	Not significantly different

Table 6: Contd.,

Numerical Data	Languages	$\left\|\frac{Q_l}{r_l}-\frac{Q_k}{r_k}\right\|$	$t_{1-(\alpha/2)}\left(D^2\frac{r-1-T}{r-s}\right)^{\frac{1}{2}}\left(\frac{1}{r_l}+\frac{1}{r_k}\right)^{\frac{1}{2}}$	Inference
(iii)Maximum number of sentences within the direct speech	1 and 2	4867.8037	2405.474	significantly different
	1and 3	4264.3312	2016.709	significantly different
	1 and 4	4121.7958	2807.14	significantly different
	2 and 3	603.4725	2090.851	Not significantly different
	2 and 4	8989.5995	2860.87	Significantly different
	3 and 4	8386.1270	2542.713	significantly different
(v) Total number of words within Direct speeches	1 and 2	1617.7185	2586.591	Not significantly different
	1and 3	3156.3645	2168.554	significantly different
	1 and 4	5926.675	3018.5	significantly different
	2 and 3	4774.0830	2248.279	significantly different
	2 and 4	7544.3935	3076.275	Significantly different
	3 and 4	2770.3105	2734.162	significantly different
(vi)Maximum number of words within the direct speech	1 and 2	5155.6611	2458.157	significantly different
	1and 3	984.0527	2060.877	Not significantly different
	1 and 4	5220.2417	2868.62	significantly different
	2 and 3	4171.6084	2136.643	significantly different
	2 and 4	10375.9028	2923.526	Significantly different
	3 and 4	6204.2941	2598.401	significantly different

6 CONCLUSIONS

Although the authors used direct speeches to express the similar emotions, among the different characters they provide a kind striking dissimilarity also corresponding to different languages, cultures and authors. For most of the statistical similarity is visible, when two Indian languages with two different authors are compared and the same is true for writing pattern of two English authors. However, it is surprising to observe **that the ratio of total number of sentences within the direct speeches to the total number of sentences in the text is similar to all the authors**. For this reason, we have to conduct the work of different authors based on the same themes, and also the work of the same author based on different themes.

REFERENCES

1. Duda Richard O., Hart Peter E. Stork David G. (2000), Pattern Classification (2nd ed.), John Willey and Sons Inc.
2. Conover W.J. (2006), Practical Nonparametric Statistics (3rd ed.), John Willey and Sons Inc.
3. Ying Wang (2003), Nonparametric tests for randomness, ECE 461 Project Report.
4. Clark, H., & Gerrig, R. (1990). Quotations as demonstrations. Language, 66, 764–805. doi:10.2307/414729
5. Groenewold, R., Bastiaanse, R., Nickels, L., & Huiskes, M. (2014a). Perceived liveliness and speech comprehensibility in aphasia: The effects of direct speech in auditory narratives. International Journal of Language and Communication Disorders. Advance online publication. doi:10.1111/1460-6984.12080

6. Groenewold, R., Bastiaanse, R., Nickels, L., Wieling, M., & Huiskes, M. (2014b). The effects of direct and indirect speech on discourse comprehension in Dutch listeners with and without aphasia. Aphasiology. 28, 862-884. Doi: 10.1080/02687038.2014.902916

7. Groenewold, R., Bastiaanse, R., Nickels, L., Wieling, M., & Huiskes, M. (2015*). The* differential effects of direct and indirect speech on discourse comprehension in Dutch and English listeners with and without aphasia**.** Aphasiology. 29(6), *685-704. Doi:10.1080/02687038.2014.977217*

8. Labov, W. (1972). Language in the city: Studies in the Black English vernacular. Philadelphia: University of Pennsylvania Press.

9. Li, C. N. (1986). Direct speech and indirect speech: A functional study. In F. Coulmas (Ed.), Direct and indirect speech. Berlin: Mouton de Gruyter.

10. Mayes, P. (1990). Quotation in spoken English. Studies in Language, 14, 325–363. doi:10.1075/ sl. 14.2.04May

11. Tannen, D. (1989). Talking voices: Repetition, dialogue, and imagery in conversational discourse. Cambridge: Cambridge University Press.

12. Wierzbicka, A. (1974). The semantics of direct and indirect discourse. Paper in Linguistics, 7, 267–307. doi:10.1080/08351817409370375

13. Sikdar M. & Sarmah P., (2017) Pattern Recognition In Language Model With Special *Reference To Children Stories, International Journal of Innovative Research and Advanced Studies (IJIRAS), Volume 4 Issue3.*